Can We Just Normalize War? *

Miles D. Williams   *Denison University*

Can we just log-normalize the size of interestate conflicts to analyze them? Doing so has been verboten among scholars interested in analyzing long-term trends in war size due to concerns that the most extreme wars are power-law distributed. This view has limited the scope of statistical tools and hypothesis tests considered suitable for analyzing the data. While a more flexible model for war size than the classic power-law formulation has been proposed, a formal

Keywords: war, power-law, log-normal

Introduction

The Power-law and Alternatives

One of the most pressing questions prior research on battle deaths seeks to address is how best to model variation in war intensity. International wars are said to follow Richardson's Law, a descriptive characterization of the fact most wars kill only a few combatants while a few are exceptionally deadly. This insight is owed to early contributions to the quantitative study of war by Lewis Fry Richardson (1948, 1960). After assembling datasets of historical wars, Richardson concluded that war size is (1) independently and identically distributed or i.i.d. and (2) follows the power-law distribution.

Datasets on historical wars have come a long way since Richardson's initial findings. In this study, the primary data source considered will be the Correlates of War (CoW) interstate conflict dataset (Sarkees and Wayman 2010), which documents the battle deaths from 95 interstate wars fought between 1816 and 2007. While some studies rely on their own original historical data, CoW is one of the most commonly used and authoritative datasets available for studying international conflict. Of course, this does not mean it is above reproach. A number of wars in the battle series kill a surprisingly round number

---

of combatants, which is more likely the result of data limitations and rounding error than highly precise killing efficiency.

Figure 1 shows the distribution of total battle deaths from the 95 wars in the dataset. For ease of interpretation, battle deaths are shown on the log-10 scale. It is plain to see that the distribution abides by Richardson's Law. In particular, the bottom 80% of wars in terms of deadliness account for only 0.38% of total battle deaths in the data. That means 20% of wars are responsible for 99.62% of battle related deaths in interstate conflicts.
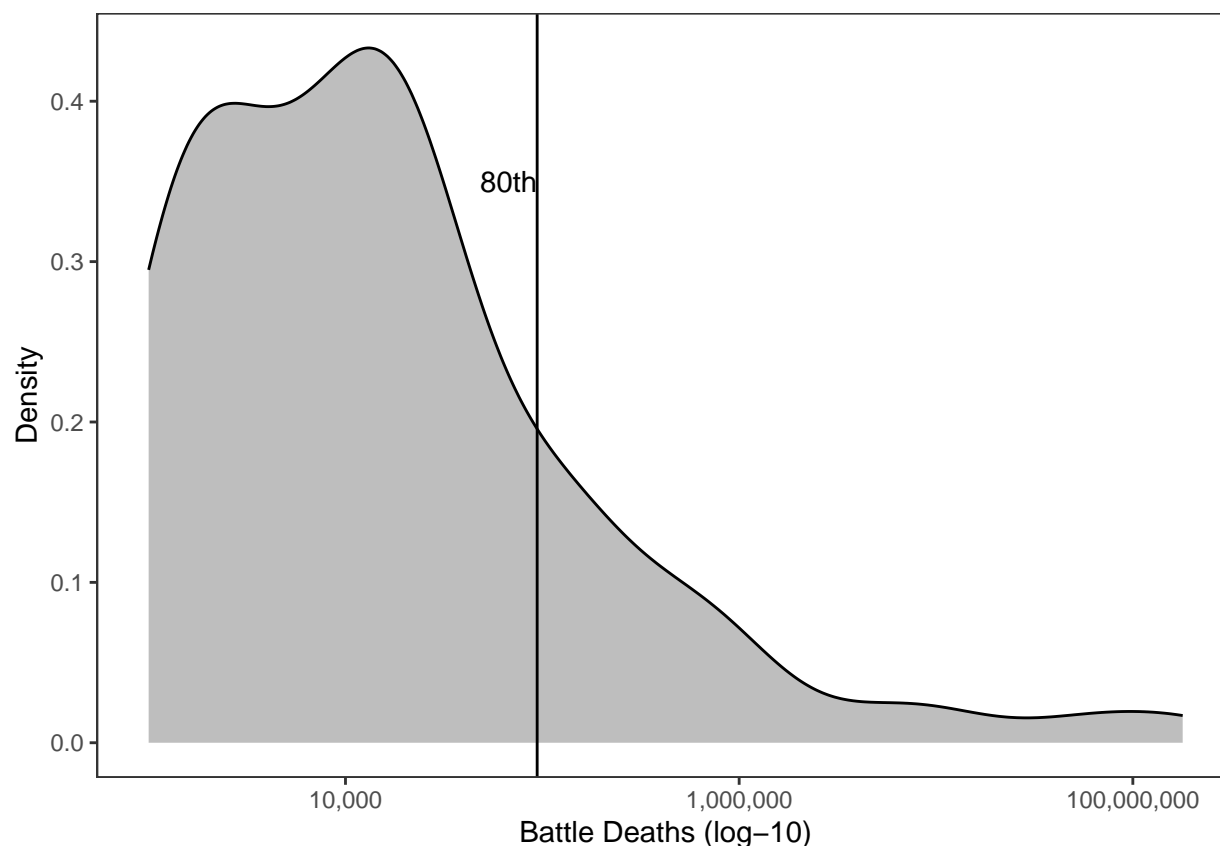

Figure 1: Density plot of total battle deaths from CoW battle series, 1816-2007. The x-axis is on the log-10 scale. The 80th percentile is denoted with a vertical line.

The power-law remains the nearly unquestioned model of choice for analyzing the distribution of battle deaths from different wars because it is generally well-suited to phenomena with extreme tails—and war size clearly has an extreme tail. The power-law model characterizes the inverse cumulative distribution function (CDF), or the probabil-

ity of an event of size X greater than x as

$$\Pr(X > x) = \propto x^{-\alpha} \quad \text{for all large } x$$

where $\alpha > 0$. That is, the probability of an event $X > x$ is inversely proportional to the size of the event raised to the power α. As $\alpha \to 0$, the likelihood of even extremely large events is quite high. This seems like an appropriate choice for studying interstate conflict, since such a small number of conflicts can account for the vast majority of battle deaths.

The power-law model also has the distinct property of linearity between the inverse CDF and observed event size on the log-log scale. That is:

$$\log[\Pr(X > x)] \propto -\alpha \log(x).$$

This is illustrated in Figure 2, which compares the theoretical inverse CDF of a hypothetical variable x on an unadjusted scale versus a log-log scale.
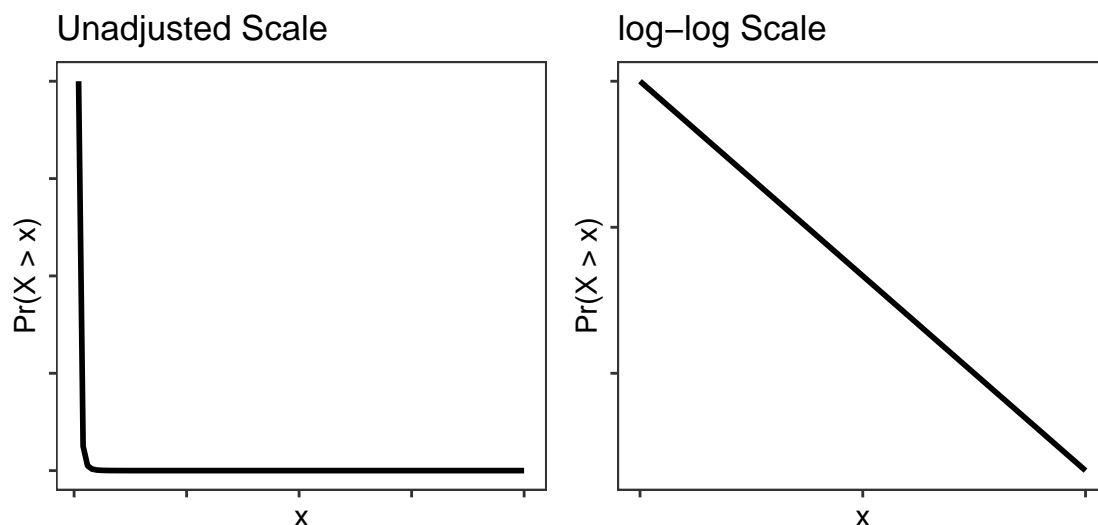

Figure 2: The inverse CDF of power-law data in unadjusted scale versus log-log scale.

In practice, such linearity is rarely so consistent across the entire set of observed data. Sometimes we only observe this characteristic linearity in the extreme tail of the distribu-

tion, giving rise to the necessity of identifying $x_{\min}$ such that all $x \geq x_{\min}$ are power-law distributed. This step of identifying $x_{\min}$, in fact, is the state-of-the-art for fitting the power-law to data (Clauset, Shalizi, and Newman 2009). The consequences can sometimes be minimal, but in other cases this approach can lead to substantial data loss. However, this can be justified if we really believe the data are power-law distributed in the extreme tail, and if such events are the primary focus of study.

Studying the extreme tails of phenomena with the power-law model, even with substantial truncation, remains a popular choice because of its mathematical properties as well. Most notable among these is the possibility of identifying scale-free phenomena. If $\alpha < 3$, the data lack a well defined variance, and if $\alpha < 2$ the data lack a well defined mean. In such extreme cases, the phenomenon under study can be subject to *black swan* behavior—events that are exceptionally extreme and inexplicable—with the expected magnitude of such events statistically indistinguishable from infinity. This has made the power-law especially relevant to conflict scholars interested in studying the deadly potential of international war. In fact, previous research on the CoW battle series finds that $\alpha < 2$, leading the author of said study to contemplate the possibility that "we're all going to die" as a result of interstate conflict (Braumoeller 2019).

Of course, the classic power-law model is not the only model that can capture data with a power-law tail. There are alternatives that have more favorable, though less provocative, properties for statistical analysis. In an excellent study, Cunen, Hjort, and Nygård (2020) recently used a more general distributional form known as the inverse Burr to study the CoW battle series summarized above. The inverse Burr distribution specifies the probability of an event greater than size x as:

$$\Pr(X > x) = 1 - \left[\frac{(x/\mu)^\theta}{1 + (x/\mu)^\theta} \right]^\alpha$$

where the parameters μ, θ, and α are strictly greater than zero. The parameter μ is a scaling

parameter that captures the central tendency of x, while θ and α are shape parameters. θ in particular functions much the same way that α does in characterizing the extreme tails of power-law distributed data. This is because as x increases we have:

$$\Pr(X > x) \approx \alpha(\mu/z)^{\theta}.$$

According to Cunen, Hjort, and Nygård (2020), the strength of the inverse Burr relative to the classic power-law is its ability to model the entire battle series, not just the extreme tail. Returning to the relationship between the inverse CDF and observed data in log-log space, the inverse Burr is able to accommodate deviations from linearity that the power-law model cannot efficiently capture. This is summarized in Figure 3, which shows the relationship between the inverse CDF and observed data assuming an inverse Burr Distribution.
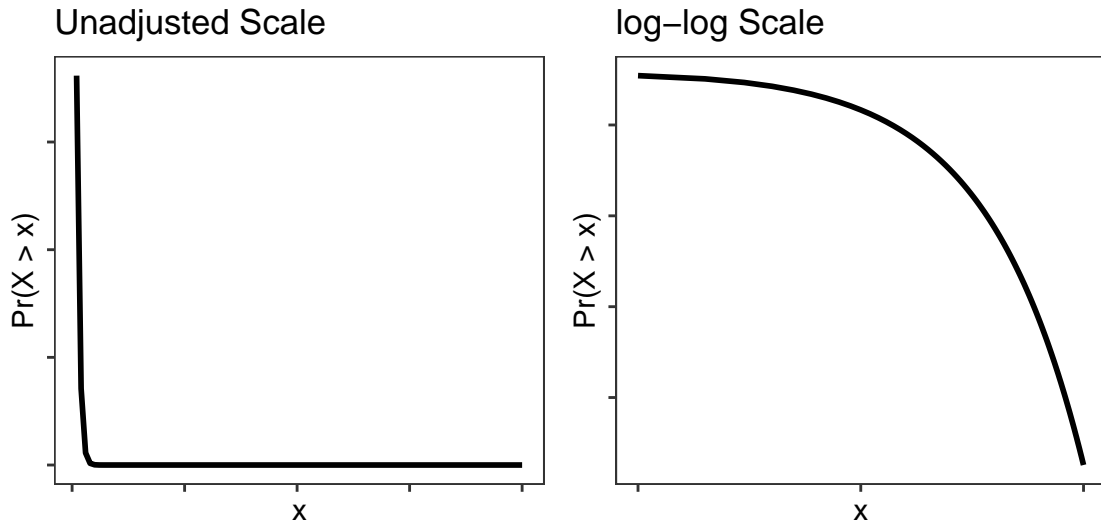

Figure 3: The inverse CDF of inverse Burr data in unadjusted scale versus log-log scale.

The result of this increased flexibility is greater statistical power because the model can be fit efficiently with all the data, not just the most extreme events. This is the justification that Cunen, Hjort, and Nygård (2020) make for preferring the inverse Burr over the classic

power-law in their analysis of the CoW battle data.

However, the inverse Burr model is not the only distributional form that can more flexibly model both smaller and larger events simultaneously, nor is it necessarily the most useful. The log-normal distribution, as the name suggests, characterizes data that is normally distributed in the log-scale. It has the inverse CDF:

$$\Pr(X > x) = 1 - \Phi([\log(x) - \mu]/\sigma)$$

where μ and $\sigma > 0$ are the logmean and log standard deviation, respectively, and $\Phi(\cdot)$ is the normal CDF.

Like the inverse Burr, the log-normal model can capture a curved relationship between the inverse CDF and the data in log-log space, as shown in Figure 4. Note that the fit is not identical to that of the inverse Burr. It is slightly more severe, which leads the log-normal model to give a slightly higher probability to smaller events and a lower probability to larger events.
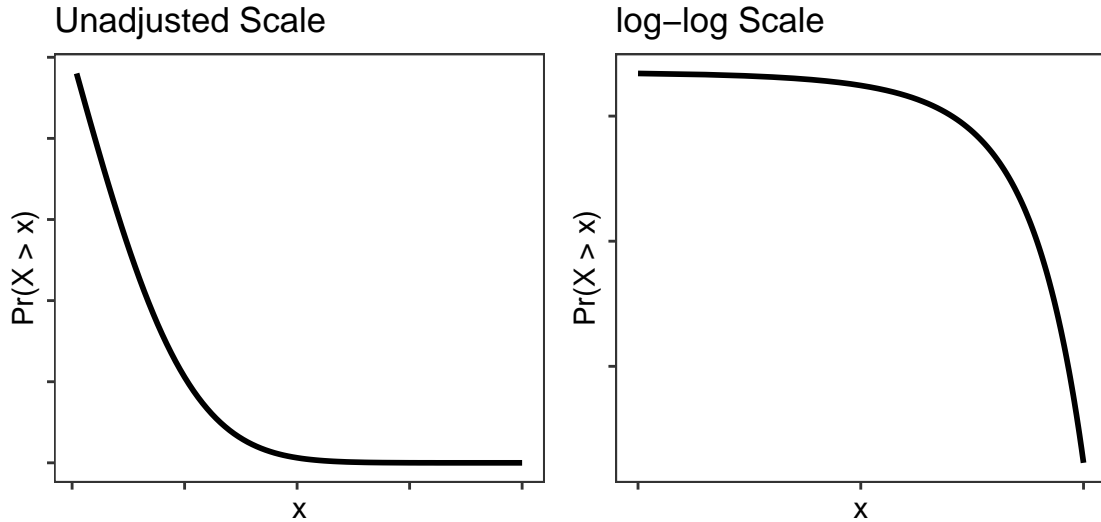

Figure 4: The inverse CDF of log-normal data in unadjusted scale versus log-log scale.

The log-normal model offers some advantages over both the classic power-law and

inverse Burr models. The first is parsimony. With one less parameter than the inverse Burr, the log-normal model has a simpler functional form. While this comes at the cost of some flexibility, a simple model that does relatively well is sometimes preferable to a more complex one that is over-fit.

The second advantage of the log-normal model is that it is amenable to conventional statistical analysis. The power-law is not straightforward to parameterize with covariates, and under some scenarios it even may be inappropriate to do so (such as with scale-free phenomena). The better behaved inverse Burr can be parameterized using covariates, but most statistical packages do not provide the option to specify an inverse Burr regression model. This means the researcher must program an inverse Burr regression model "by hand." Meanwhile, log-normal data can be modeled using conventional regression analysis tools with model parameters estimated either via ordinary least squares or maximum likelihood.

For practical reasons, if the log-normal model makes for a suitable fit for interstate war size, this opens a world of possibility for making new inferences and testing new hypotheses with historical war data. Even so, past research on war size has typically overlooked or minimized the utility of log-normalization. One of the contributions of this study is to provide evidence-based assessments of the appropriateness of the log-normal model for studying war size.

Methods for Model Fitting and Selection

Clauset, Shalizi, and Newman (2009) provide tools for using the power-law to study phenomena, test its goodness of fit, and compare it to alternatives. However, while the statistical software they developed supports the power-law and log-normal distributions, it does not support the inverse Burr. The code necessary to do so had to be written from scratch, which is provided in the Code Appendix.

Clauset, Shalizi, and Newman (2009) lay out a simple "recipe" for analyzing data

with the power-law model. The first step is simply to fit the power-law to data using the procedure summarized in the previous section. In short, find the best fitting $x_{\min}$ and $\hat{\alpha}$.

The second step in the recipe is to perform a goodness of fit (GOF) test. Clauset, Shalizi, and Newman (2009) detail a bootstrapping/simulation approach that involves first estimating the KS statistic for the fitted power-law model. Next, using the fitted model, simulate new synthetic datasets. The power-law is then fit to the synthetic data and a KS statistic is calculated using the synthetic data and the power-law parameters fit to that specific synthetic dataset. A p-value from this test is calculated as the share of times the distances measured using the simulated data are larger than the empirical distance. If the p-value is small then the power-law is not a plausible fit for the data. Clauset, Shalizi, and Newman (2009) recommend $p < 0.1$ as the cutoff.

The third and final step involves comparing the power-law to alternative models. One way to approach this is to fit the alternatives to the same data, perform the GOF test proposed above but with the alternative models, and compare the results to the power-law. If the power-law cannot be ruled out, but the alternatives can, this serves as a basis for favoring the power-law.

However, in some cases separate GOF tests will not be definitive. In this case, Clauset, Shalizi, and Newman (2009) suggest a likelihood ratio (LR) test to formally judge which model is the better fit for the data. In particular, they recommend Vuong's test, which is an LR-based test for model selection using Kullback-Leibler criteria (see **vuong1989likelihood?**).

Using this simple recipe, it is not possible to identify the "true" model for the data. However, it provides a rigorous and data-driven way to identify the best model for the data. The next section outlines the data sources and measures that will be used to identify the best-fitting model for battle deaths.

Data and Measures

As noted earlier, the data used to fit the models comes from the CoW interstate conflict dataset, which documents battle deaths per country across 95 interstate wars fought between 1816 and 2007. The data were accessed using the {peacesciencer} R package (Miller 2022). For each conflict in the dataset, the total number of battle deaths across countries fighting a war were tallied, yielding a battle deaths series of 95 observations.

As noted earlier, other datasets have been assembled of historical wars. [cite some of them here] The CoW conflict series is chosen for two reasons. The first is its widespread use, which makes comparisons with approaches used in other studies easier [cites here]. The second justification is that the CoW data are generally considered of good quality. That does not mean they are regarded as perfect. As previously mentioned, the data contain a number of suspiciously round numbers for total battle deaths, meaning the counts likely are not perfectly accurate for several conflicts. There are also some conflicts in the data for which death counts are disputed [cite here]. The data also have limited coverage. While the data cover the years 1816 to 2007 (nearly two centuries worth of wars), the history of human conflict did not start in 1816, nor did it end in 2007. Some studies have developed conflict series that start at year 1 A.C.E. and run to the present [cite]. However, the quality of these data are less established.

With a dataset chosen, the next issue to settle is how to measure war size. Should total battle deaths be studied, or should deaths be normalized by population? The answer depends on what question we want to answer. Total battle deaths provide a direct measure of absolute war size, while relative quantities, like deaths per 1 million, provide a measure of relative risk of dying in war. Different studies take different approaches. The choice is made here to apply both, since both quantities can be of interest depending on the research questions being asked. It is also possible that normalizing by population will alter conclusions about model fit. It could be that the power-law is appropriate for

modeling the absolute size of wars, while the inverse Burr or log-normal are better once population size is accounted for.

With respect to relative conflict size, the choice of denominator also has to be addressed. Some scholars prefer to use global population as the denominator rather than only the populations of the countries involved in a war [cite]. This choice follows from different goals. Deaths per global population is akin to a measure of all-cause mortality from war, meaning it treats war like a public health issue. Deaths per the populations of the countries fighting a war treats war like a political activity or behavior that has unique consequences for the countries involved (Braumoeller 2019). Rather than adjudicate between which approach is most appropriate, both measures, in addition to total war size, are utilized here.

The dataset to be analyzed, then, has for each given war i in $i = 1, ..., 95$ a measure of absolute war size (x_i), a measure of war size per global population (x_i/n_i), and a measure of war size per the populations of the countries fighting a given war (x_i/m_i). The value n_i denotes world population at the time a war starts, and m_i denotes the combined population of the countries fighting a war at the time a war starts. Values for each were scaled to deaths per 1 million. The data for population come from version 6.0 of the CoW National Military Capabilities dataset, and were also accessed using the {peacesciencer} package (D. J. Singer 1987; J. D. Singer, Bremer, and Stuckey 1972).

Analysis

Having established the conflict series and measures of war size to be used for the analysis, the next step is to fit the alternative models to the data following the "recipe" outlined by Clauset, Shalizi, and Newman (2009) for evaluating their performance. In the first section below, a descriptive summary of each of the model fits for the data is provided. The analysis was supported by software developed by Clauset, Shalizi, and Newman (2009) for estimating and validating power-law fits for data relative to alternative distributions.

Table 1: GOF tests for total war size

Model	GOF	p-value
Power-law	0.07	0.85
Inverse Burr	0.12	0.01
Log-normal	0.08	0.05

Table 2: GOF tests for war size per global population

Model	GOF	p-value
Power-law	0.05	0.94
Inverse Burr	0.08	0.23
Log-normal	0.09	0.04

Unfortunately, the software does not currently support the inverse Burr distribution, but modifications to the code allow capability for the inverse Burr to be grafted in. The R code is available in the code Appendix for interested readers.
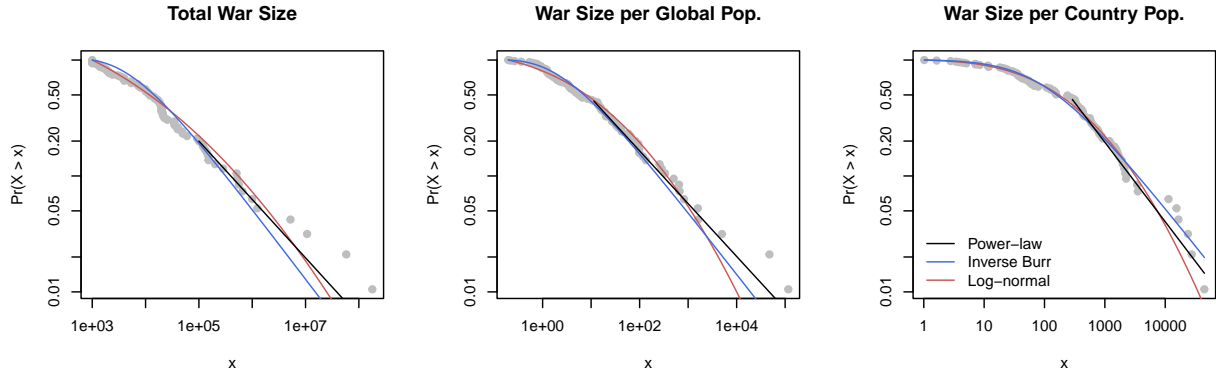

Figure 5: Visualization of model fits for the data. The empirical inverse CDF is shown over the data in log-log space across panels. The first panel shows results for total battle deaths. The second shows results for battle deaths per global population in millions. The third shows results for battle deaths per population of the countries at war in millions.

Table 3: GOF tests for war size per country population

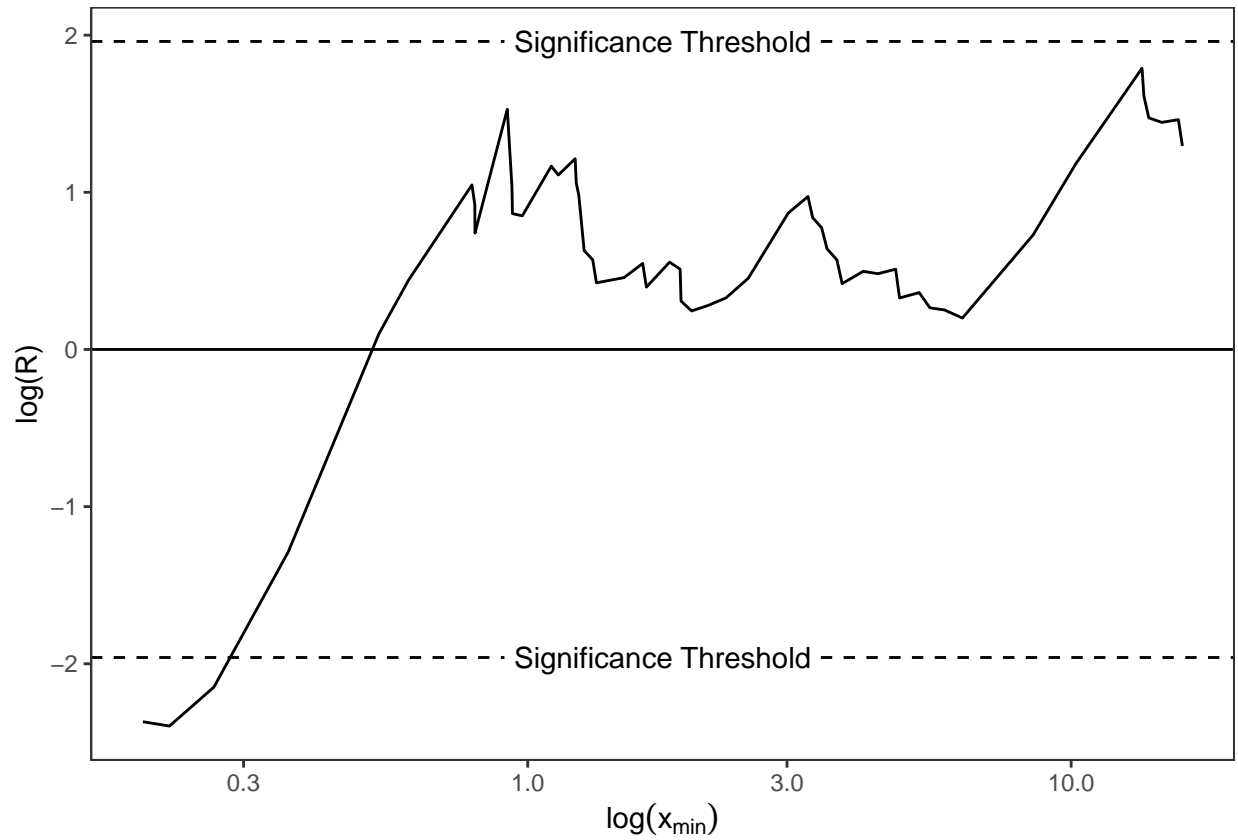| Model	GOF	p-value
Power-law	0.06	0.84
Inverse Burr	0.08	0.09
Log-normal	0.06	0.35

Figure 6: Likelihood ratio test for power-law vs. inverse Burr model using battle deaths per global population in millions. Values shown over possible data trunctions.
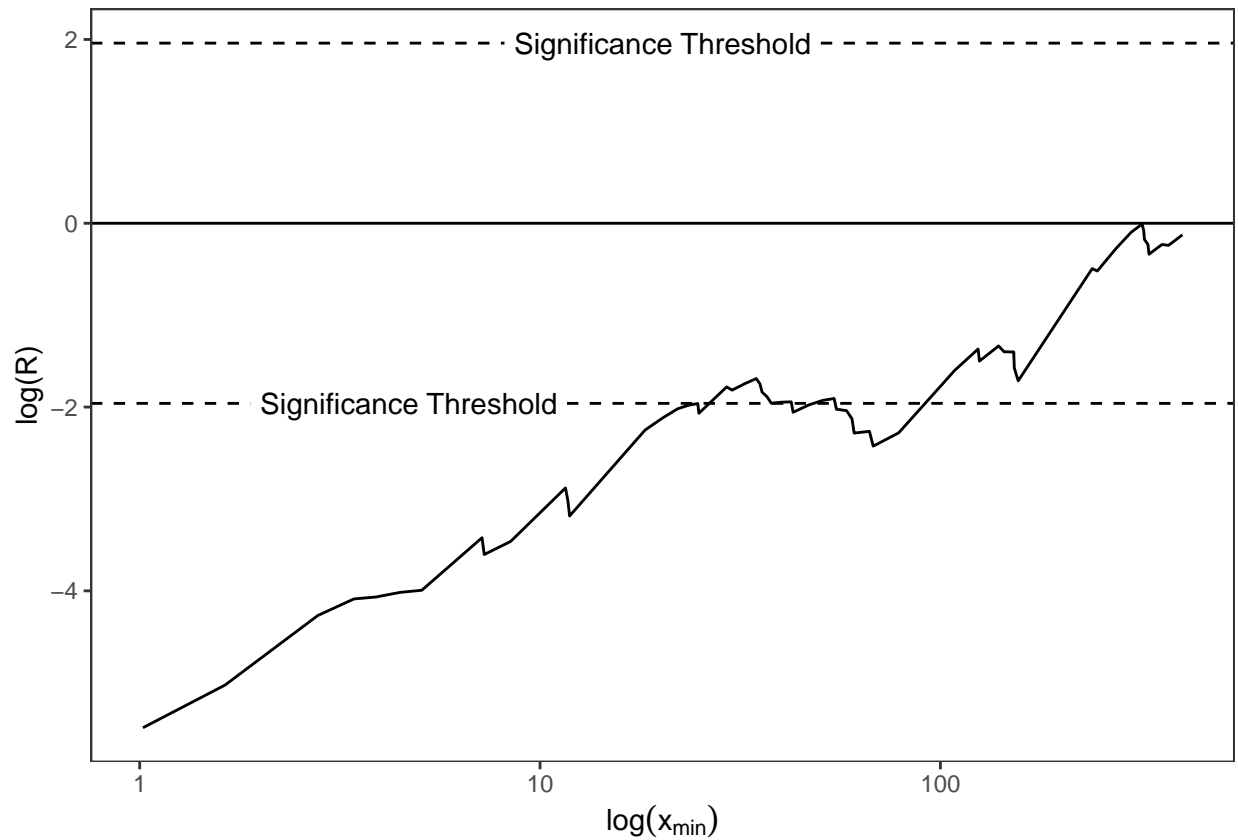
Figure 7: Likelihood ratio test for power-law vs. log-normal model using battle deaths per country population in millions. Values shown over possible data trunctions.

Table 4: Is 1950 a significant change point for war size?

	Diff.	SE	Stat	p-value
α	0.02	0.53	0.03	0.97
μ-log	-1.11	0.42	-2.63	0.01

Comparison of model fits for pre- and post-1950 battle deaths per country populations in millions. The difference in power-law slopes is shown in the first row. The difference in the mean-log is shown in the second row.

Implications and Recommendations

References

Braumoeller, Bear F. 2019. *Only the Dead: The Persistence of War in the Modern Age*. New York: Oxford University Press.

Clauset, Aaron, Cosma Rohilla Shalizi, and M.E.J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51 (4): 661–703.

Cunen, Céline, Nils Lid Hjort, and Håvard Mokleiv Nygård. 2020. "Statistical Sightings of Better Angels: Analysing the Distribution of Battle-Deaths in Interstate Conflict over Time." *Journal of Peace Research* 57 (2): 221–34.

Miller, Steven V. 2022. "peacesciencer: An r Package for Quantitative Peace Science Research." *Conflict Management and Peace Science*. http://svmiller.com/peacesciencer/.

Richardson, Lewis F. 1948. "Variation of the Frequency of Fatal Quarels with Magnitude." *American Statistical Association* 43: 523–46.

———. 1960. *Statistics of Deadly Quarels*. Chicago: Quadrangle Books.

Sarkees, Meredith Reid, and Frank Wayman. 2010. *Resort to War: 1816 - 2007*. Washington DC: CQ Press.

Singer, David J. 1987. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985." *International Interactions* 14 (1): 115–32.

Singer, J. David, Stuart A. Bremer, and John Stuckey. 1972. "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." In *Peace, War and Numbers*, edited by Bruce Russett. Beverly Hills, CA: Sage Publications, Inc.